



---

[Replication and Meta-Analysis in Parapsychology]: Rejoinder

Author(s): Jessica Utts

Source: *Statistical Science*, Vol. 6, No. 4 (Nov., 1991), pp. 396-403

Published by: [Institute of Mathematical Statistics](#)

Stable URL: <http://www.jstor.org/stable/2245736>

Accessed: 01/07/2014 03:36

---

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at  
<http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



*Institute of Mathematical Statistics* is collaborating with JSTOR to digitize, preserve and extend access to *Statistical Science*.

<http://www.jstor.org>

after a careful analysis is completed, there can be vigorous reasonable arguments about the appropriateness of the formulation and its analysis. These investigations leave me reinforced with the belief that people cannot do hard mathematical problems in their heads, rather than with an attitude toward or against ESP investigations.

When I first became aware of the work of Rhine and others, the concept seemed to me to be very important and I asked a psychologist friend why more psychologists didn't study this field. He responded that there were too many ways to do these experiments in a poorly controlled manner. At the time, I had just discovered that when viewed with light coming from a certain angle, I could read the

backs of the cards of my parapsychology deck as clearly as the faces. While preparing these remarks in 1991, I found a note on page 305 of volume 1 of *The Journal of Parapsychology* (1937) indicating that imperfections in the cards precluded their use in unscreened situations, but that improvements were on the way. Thus I sympathize with Utts's conclusion that much is to be gained by studying how to carry out such work well. If there is no ESP, then we want to be able to carry out null experiments and get no effect, otherwise we cannot put much belief in work on small effects in non-ESP situations. If there is ESP, that is exciting. However, thus far it does not look as if it will replace the telephone.

## Rejoinder

Jessica Utts

I would like to thank this distinguished group of discussants for their thought-provoking contributions. They have raised many interesting and diverse issues. Certain points, such as Professor Mosteller's enlightening account of Feller's position, require no further comment. Other points indicate the need for clarification and elaboration of my original material. Issues raised by Professors Diaconis and Hyman and subsequent conversations with Robert Rosenthal and Charles Honorton have led me to consider the topic of "Satisfying the Skeptics." Since the conclusion in my paper was not that psychic phenomena have been proved, but rather that there is an anomalous effect that needs to be explained, comments by several of the discussants led me to address the question "Should Psi Research be Ignored by the Scientific Community?" Finally, each of the discussants addressed replication and modeling issues. The last part of my rejoinder comments on some of these ideas and discusses them in the context of parapsychology.

### CLARIFICATION AND ELABORATION

Since my paper was a survey of hundreds of experiments and many published reports, I could obviously not provide all of the details to accompany this overview. However, there were details lacking in my paper that have led to legitimate questions and misunderstandings from several of the discussants. In this section, I address specific points raised by Professors Diaconis, Greenhouse,

Hyman and Morris, by either clarifying my original statements or by adding more information from the original reports.

### Points Raised by Diaconis

Diaconis raised the point that qualified skeptics and magicians should be active participants in parapsychology experiments. I will discuss this general concept in the next section, but elaborate here on the steps that were taken in this regard for the autoganzfeld experiments described in Section 5 of my paper. As reported by Honorton et al. (1990):

Two experts on the simulation of psi ability have examined the autoganzfeld system and protocol. Ford Kross has been a professional mentalist [a magician who simulates psychic abilities] for over 20 years... Mr. Kross has provided us with the following statement: "In my professional capacity as a mentalist, I have reviewed Psychophysical Research Laboratories' automated ganzfeld system and found it to provide excellent security against deception by subjects." We have received similar comments from Daryl Bem, Professor of Psychology at Cornell University. Professor Bem is well known for his research in social and personality psychology. He is also a member of the Psychic Entertainers Association and has performed for many years as a mentalist. He vis-

ited PRL for several days and was a subject in Series 101" [pages 134–135].

Honorton has also informed me (personal communication, July 25, 1991) that several self-proclaimed skeptics have visited his laboratory and received demonstrations of the autoganzfeld procedure and that no one expressed any concern with the security arrangements.

This may not completely satisfy Professor Diaconis' objections, but it does indicate a serious effort on the part of the researchers to involve such people. Further, the original publication of the research in Section 5 followed the reporting criteria established by Hyman and Honorton (1986), thus providing much more detail for the reader than the earlier published records to which Professor Diaconis alludes.

### Points Raised by Greenhouse

Greenhouse enumerated four items that offer alternative explanations for the observed anomalous effects. Three of these (items 2–4) will be addressed in this section by elaborating on the details provided in my paper. His item 1 will be addressed in a later section.

Item 2 on his list questioned the role of experimenter expectancy effects as a potential confounder in parapsychological research. While the expectations of the experimenter may influence the *reporting* of results, the ganzfeld experiments (as well as other psi experiments) are conducted in such a way that experimenter expectancy cannot account for the results themselves. Rosenthal, who Greenhouse cites as the expert in this area, addressed this in his background paper for the National Research Council (Harris and Rosenthal, 1988a) and concluded that the ganzfeld studies were adequately controlled in this regard. He also visited the autoganzfeld laboratory and was given a demonstration of that procedure.

Greenhouse's item 3, the question of what constitutes a direct hit, was addressed in my paper but perhaps needs elaboration. Although free-response experiments do generate substantial amounts of subjective data, the statistical analysis requires that the results for each trial be condensed into a single measure of whether or not a direct hit was achieved. This is done by presenting four choices to a judge (who of course does not know the correct answer) and asking the judge to decide which of the four best matches the subject's response. If the judge picks the target, a direct hit has occurred.

It is true that different judges may differ on their opinions of whether or not there has been a direct hit on any given trial, but in all cases the statisti-

cal question is the same. Under the null hypothesis, since the target is randomly selected from the four possibilities presented, the probability of a direct hit is 0.25 regardless of who does the judging. Thus, the observed anomalous effects cannot be explained by assuming there was an over-optimistic judge.

If Professor Greenhouse is suggesting that the source of judging may be a moderating variable that determines the magnitude of the demonstrated anomalous effect, I agree. The parapsychologists have considered this issue in the context of whether or not subjects should serve as judges for their own sessions, with differing opinions in different laboratories. This is an example of an area that has been suggested for further research.

Finally, Greenhouse raised the question of the accuracy of the file-drawer estimates used in the reported meta-analyses. I agree that it is instructive to examine the file-drawer estimate using more than one model. As an example, consider the 39 studies from the direct hit and autoganzfeld data bases. Rosenthal's fail-safe  $N$  estimates that there would have to be 371 studies in the file-drawer to account for the results. In contrast, the method proposed by Iyengar and Greenhouse gives a file-drawer estimate of 258 studies. Even this estimate is unrealistically large for a discipline with as few researchers as parapsychology. Given that the average number of trials per experiment is 30, this would represent almost 8000 unreported trials, and at least that many hours of work.

There are pros and cons to any method of estimating the number of unreported studies, and the actual practices of the discipline in question should be taken into account. Recognizing publication bias as an issue, the Parapsychological Association has had an official policy since 1975 against the selective reporting of positive results. Of the original ganzfeld studies reported in Section 4 of my paper, less than half were significant, and it is a matter of record that there are many nonsignificant studies and "failed replications" published in all areas of psi research. Further, the autoganzfeld database reported in Section 5 has no file-drawer. Given the publication practices and the size of the field, the proposed file-drawer cannot account for the observed effects.

### Points Raised by Hyman

One of my goals in writing this paper was to present a fair account of recent work and debate in parapsychology. Thus, I was disturbed that Hyman, who has devoted much of his career to the study of parapsychology, and who had first-hand knowledge of the original published reports, be-

lieved that some of my statements were inaccurate and indicated that I had not carefully read the reports. I will address some of his specific objections and show that, except where noted, the accuracy of my original statements can be verified by further elaboration and clarification, with due apology for whatever necessary details were lacking in my original report.

Most of our points of disagreement concern the National Academy of Sciences (National Research Council) report *Enhancing Human Performance* (Druckman and Swets, 1988). This report evaluated several controversial areas, including parapsychology. Professor Hyman chaired the Parapsychology Subcommittee. Several background papers were commissioned to accompany this report, available from the "Publication on Demand Program" of the National Academy Press. One of the papers was written by Harris and Rosenthal, and entitled "Human Performance Research: An Overview."

Professor Hyman alleged that "Utts mistakenly asserts that my subcommittee on parapsychology commissioned Harris and Rosenthal to evaluate parapsychology experiments for us . . ." I cannot find a statement in my paper that asserts that Harris and Rosenthal were commissioned by the subcommittee, nor can I find a statement that asserts that they were asked to evaluate parapsychology experiments. Nonetheless, I believe our substantive disagreement results from the fact that the work by Harris and Rosenthal was written in two parts, both of which I referenced in my paper. They were written several months apart, but published together, and each had its own history.

The first part (Harris and Rosenthal, 1988a) is the one to which I referred with the words "Rosenthal was commissioned by the National Academy of Sciences to prepare a background paper to accompany its 1988 report on parapsychology" (p. 372). According to Rosenthal (personal communication, July 23, 1991) he was asked to prepare a background paper to address evaluation issues and experimenter effects to accompany the report in five specific areas of research, including parapsychology.

The second part was a "Postscript" to the commissioned paper (Harris and Rosenthal, 1988b), and this is the one to which I referred on page 371 as "requested by Hyman in his capacity as Chair of the National Academy of Sciences' Subcommittee on Parapsychology." (It is probably this wording that led Professor Hyman to his erroneous allegation.) The postscript began with the words "We have been asked to respond to a letter from Ray

Hyman, chair of the subcommittee on parapsychology, in which he raises questions about the presence and consequence of methodological flaws in the ganzfeld studies . . ."

In reference to this postscript, I stand corrected on a technical point, because Hyman himself did not request the response to his own letter. As noted by Palmer, Honorton and Utts (1989), the postscript was added because:

At one stage of the process, John Swets, Chair of the Committee, actually phoned Rosenthal and asked him to withdraw the parapsychology section of his [commissioned] paper. When Rosenthal declined, Swets and Druckman then requested that Rosenthal respond to criticisms that Hyman had included in a July 30, 1987 letter to Rosenthal [page 38].

A related issue on which I would like to elaborate concerns the correlation between flaws and success in the original ganzfeld data base. Hyman has misunderstood both my position and that of Harris and Rosenthal. He believes that I implicitly denied the importance of the flaws, so I will make my position explicit. I do not think there is any evidence that the experimental results were due to the identified flaws. The flaw analysis was clearly useful for delineating acceptable criteria for future experiments. Several experiments were conducted using those criteria. The results were similar to the original experiments. I believe that this indicates an anomaly in need of an explanation.

In discussing the paper and postscript by Harris and Rosenthal, Hyman stated that "The alleged contradictory conclusions [to the National Research Council report] of Harris and Rosenthal are based on a meta-analysis that supports Honorton's position when Honorton's [flaw] ratings are used and supports my position when my ratings are used." He believes that Harris and Rosenthal (and I) failed to see this point because the low power of the test associated with their analysis was not taken into account.

The analysis in question was based on a canonical correlation between flaw ratings and measures of successful outcome for the ganzfeld studies. The canonical correlation was 0.46, a value Hyman finds to be impressive. What he has failed to take into account however, is that a canonical correlation gives only the *magnitude* of the relationship, and not the *direction*. A careful reading of Harris and Rosenthal (1988b) reveals that their analysis actually *contradicted* the idea that the flaws could account for the successful ganzfeld results, since "Interestingly, three of the six flaw variables correlated positively with the flaw canonical variable



and with the outcome canonical variable but three correlated *negatively*" (page 2, italics added). Rosenthal (personal communication, July 23, 1991) verified that this was indeed the point he was trying to make. Readers who are interested in drawing their own conclusions from first-hand analyses can find Hyman's original flaw codings in an Appendix to his paper (Hyman, 1985, pages 44–49).

Finally, in my paper, I stated that the parapsychology chapter of the National Research Council report critically evaluated statistically significant experiments, but not those that were nonsignificant. Professor Hyman "does not know how [I] got such an impression," so I will clarify by outlining some of the material reviewed in that report. There were surveys of three major areas of psi research: remote viewing (a particular type of free-response experiment), experiments with random number generators, and the ganzfeld experiments. As an example of where I got the impression that they evaluated only significant studies, consider the section on remote viewing. It began by referencing a published list of 28 studies. Fifteen of these were immediately discounted, since "only 13... were published under refereed auspices" (Druckman and Swets, 1988, page 179). Four more were then dismissed, since "Of the 13 scientifically reported experiments, 9 are classified as successful" (page 179). The report continued by discussing these nine experiments, never again mentioning any of the remaining 19 studies. The other sections of the report placed similar emphasis on significant studies. I did not think this was a valid statistical method for surveying a large body of research.

#### Minor Point Raised by Morris

The final clarification I would like to offer concerns the minor point raised by Professor Morris, that "When Honorton omitted studies that did not report direct hits as a measure, he may have biased his sample." This possibility was explicitly addressed by Honorton (1985, page 59). He examined what would happen if *z*-scores of zero were inserted for the 10 studies for which the number of direct hits was not measured, but could have been. He found that even with this conservative scenario, the combined *z*-score only dropped from 6.60 to 5.67.

#### SATISFYING THE SKEPTICS

Parapsychology is probably the only scientific discipline for which there is an organization of skeptics trying to discredit its work. The Committee for the Scientific Investigation of Claims of the

Paranormal (CSICOP) was established in 1976 by philosopher Paul Kurtz and sociologist Marcello Truzzi when "Kurtz became convinced that the time was ripe for a more active crusade against parapsychology and other pseudo-scientists" (Pinch and Collins, 1984, page 527). Truzzi resigned from the organization the next year (as did Professor Diaconis) "because of what he saw as the growing danger of the committee's excessive negative zeal at the expense of responsible scholarship" (Collins and Pinch, 1982, page 84). In an advertising brochure for their publication *The Skeptical Inquirer*, CSICOP made clear its belief that paranormal phenomena are worthy of scientific attention only to the extent that scientists can fight the growing interest in them. Part of the text of the brochure read: "Why the sudden explosion of interest, even among some otherwise sensible people, in all sorts of paranormal 'happenings'?... Ten years ago, scientists started to fight back. They set up an organization—The Committee for the Scientific Investigation of Claims of the Paranormal."

During the six years that I have been working with parapsychologists, they have repeatedly expressed their frustration with the unwillingness of the skeptics to specify what would constitute acceptable evidence, or even to delineate criteria for an acceptable experiment. The Hyman and Honorton Joint Communiqué was seen as the first major step in that direction, especially since Hyman was the Chair of the Parapsychology Subcommittee of CSICOP.

Hyman and Honorton (1986) devoted eight pages to "Recommendations for Future Psi Experiments," carefully outlining details for how the experiments should be conducted and reported. Honorton and his colleagues then conducted several hundred trials using these specific criteria and found essentially the same effect sizes as in earlier work for both the overall effect and effects with moderator variables taken into account. I would expect Professor Hyman to be very interested in the results of these experiments he helped to create. While he did acknowledge that they "have produced intriguing results," it is both surprising and disappointing that he spent only a scant two paragraphs at the end of his discussion on these results.

Instead, Hyman seems to be proposing yet another set of requirements to be satisfied before parapsychology should be taken seriously. It is difficult to sort out what those requirements should be from his account: "[They should] specify, in advance, the complete sample space and the critical region. When they get to the point where they can specify this along with some boundary conditions and make some reasonable predictions, then they

will have demonstrated something worthy of our attention."

Diaconis believes that psi experiments do not deserve serious attention unless they actively involve skeptics. Presumably, he is concerned with subject or experimenter fraud, or with improperly controlled experiments. There are numerous documented cases of fraud and trickery in purported psychic phenomena. Some of these were observed by Diaconis and reported in his article in *Science*. Such cases have mainly been revealed when investigators attempted to verify the claims of individual psychic practitioners in quasi-experimental or uncontrolled conditions. These instances have received considerable attention, probably because the claims are so sensational, the fraud is so easy to detect by a skilled observer and they are an easy target for skeptics looking for a way to discredit psychic phenomena. As noted by Hansen (1990), "Parapsychology has long been tainted by the fraudulent behavior of a few of those claiming psychic abilities" (page 25).

Control against deception by subjects in the laboratory has been discussed extensively in the parapsychological literature (see, e.g., Morris, 1986, and Hansen, 1990). Properly designed experiments should preclude the possibility of such fraud. Hyman and Honorton (1986, page 355) explicitly discussed precautions to be taken in the ganzfeld experiments, all of which were followed in the autoganzfeld experiments. Further the controlled laboratory experiments discussed in my paper usually used a large number of subjects, a situation that minimizes the possibility that the results were due to fraud on the part of a few subjects. As for the possibility of experimenter fraud, it is of course an issue in all areas of science. There have been a few such instances in parapsychology, but since parapsychologists tend to be aware of this possibility, they were generally detected and exposed by insiders in the field.

It is not clear whether or not Diaconis is suggesting that a magician or "qualified skeptic" needs to be present at all times during a laboratory experiment. I believe that it would be more productive for such consultation to occur during the design phase, and during the implementation of some pilot sessions. This is essentially what was done for the autoganzfeld experiments, in which Professor Hyman, a skeptic as well as an accomplished magician, participated in the specification of design criteria, and mentalists Bem and Kross observed experimental sessions. Bem is also a well-respected experimental psychologist.

While I believe that the skeptics, particularly some of the more knowledgeable members of

CSICOP, have served a useful role in helping to improve experiments, their counter-advocacy stance is counterproductive. If they are truly interested in resolving the question of whether or not psi abilities exist, I would expect them to encourage evaluation and experimentation by unbiased, skilled experimenters. Instead, they seem to be trying to discourage such interest by providing a moving target of requirements that must be satisfied first.

### SHOULD PSI RESEARCH BE IGNORED BY THE SCIENTIFIC COMMUNITY?

In the conclusion of my paper, I argued that the scientific community should pay more attention to the experimental results in parapsychology. I was not suggesting that the accumulated evidence constitutes proof of psi abilities, but rather that it indicates that there is indeed an anomalous effect that needs an explanation. Greenhouse noted that my paper will not necessarily change anyone's view about the existence of paranormal phenomena, an observation with which I agree. However, I hope it will change some views about the importance of further investigation.

Mosteller and Diaconis both acknowledged that there are reasons for statisticians to be interested in studying the anomalous effects, regardless of whether or not psi is real. As noted by Mosteller, "If there is no ESP, then we want to be able to carry out null experiments and get no effect, otherwise we cannot put much belief in work on small effects in non-ESP situations." Diaconis concluded that "Parapsychology is worthy of serious study" partly because "If it is wrong, it offers a truly alarming massive case study of how statistics can mislead and be misused."

Greenhouse noted several sociological reasons for the resistance of the scientific community to accepting parapsychological phenomena. One of these is that they directly contradict the laws of physics. However, this assertion is not uniformly accepted by physicists (see, e.g., Oteri, 1975), and some of the leading parapsychological researchers hold Ph.D.s in physics.

Another reason cited by Greenhouse, and supported by Hyman, is that psychic phenomena are currently unexplainable by a unified scientific theory. But that is precisely the reason for more intensive investigation. The history of science and medicine is replete with examples where empirical departures from expectation led to important findings or theoretical models. For example, the causal connection between cigarette smoking and lung cancer was established only after years of statisti-

cal studies, resulting from the observation by one physician that his lung cancer patients who smoked did not recover at the same rate as those who did not. There are many medications in common use for which there is still no medical explanation for their observed therapeutic effectiveness, but that does not prohibit their use.

There are also examples where a coherent theory of a phenomenon was impossible because the requisite background information was missing. For instance, the current theory of endorphins as an explanation for the success of acupuncture would have been impossible before the discovery of endorphins in the 1970s.

Mosteller's observation that ESP will not replace the telephone leads to the question of whether or not psi abilities are of any use even if they do exist, since the effects are relatively small. Again, a look at history is instructive. For example, in 1938 *Fortune Magazine* reported that "At present, few scientists foresee any serious or practical use for atomic energy."

Greenhouse implied that I think parapsychology is not accepted by more of the scientific community only because they have not examined the data, but this misses the main point I was trying to make. The point is that individual scientists are willing to express an opinion without any reference to data. The interesting sociological question is why they are so resistant to examining the data. One of the major reasons is undoubtedly the perception identified by Greenhouse that there is some connection between parapsychology and the occult, or worse, religious beliefs. Since religion is clearly not in the realm of science, the very thought that parapsychology might be a science leads to what psychologists call "cognitive dissonance." As noted by Griffin (1988), "People feel unpleasantly aroused when two cognitions are dissonant—when they contradict one another" (page 33). Griffin continued by observing that there are also external reasons for scientists to discount the evidence, since "It is generally easier to be a skeptic in the face of novel evidence; skeptics may be overly conservative, but they are rarely held up to ridicule" (page 34).

In summary, while it may be safer and more consonant with their beliefs for individual scientists to ignore the observed anomalous effects, the scientific community should be concerned with finding an explanation. The explanations proposed by Greenhouse and others are simply not tenable.

## REPLICATION AND MODELING

Parapsychology is one of the few areas where a point null hypothesis makes some sense. We can

specify what should happen if there is no such thing as ESP by using simple binomial models, either to find  $p$ -values or Bayes factors. As noted by Mosteller, if there is no ESP, or other nonstatistical explanation for an effect, we should be able to carry out null experiments and get no effect. Otherwise, we should be worried about using these simple models for other applications.

Greenhouse, in his first alternative explanation for the results, questioned the use of these simple models, but his criticisms do not seem relevant to the experiments discussed in Section 5 of my paper. The experiments to which he referred were either poorly controlled, in which case no statistical analysis could be valid, or were specifically designed to incorporate trial by trial feedback in such a way that the analysis needed to account for the added information. Models and analyses for such experiments can be found in the references given at the end of Diaconis' discussion.

For the remainder of this discussion, I will confine myself to models appropriate for experiments such as the autoganzfeld described in Section 5. It is this scenario for which Bayarri and Berger computed Bayes factors, and for which Dawson discussed possible Bayesian models.

If ESP does exist, it is undoubtedly a gross oversimplification to use a simple non-null binomial model for these experiments. In addition to potential differences in ability among subjects, there were also observed differences due to dynamic versus static targets, whether or not the sender was a friend, and how the receiver scored on measures of extraversion. All of these differences were anticipated in advance and could be incorporated into models as covariates.

It is nonetheless instructive to examine the Bayes factor computed by Bayarri and Berger for the simple non-null binomial model. First, the observed anomalous effects would be less interesting if the Bayes factor was small for reasonable values of  $r$ , as it was for the random number generator experiments analyzed by Jefferys (1990), most of which purported to measure psychokinesis instead of ESP. Second, the Bayes factor provides a rough measure of the strength of the evidence against the null hypothesis and is a much more sensible summary than the  $p$ -value. The Bayes factors provided by Bayarri and Berger are probably more conservative, in the sense of favoring the null hypothesis, than those that would result from priors elicited from parapsychologists, but are probably reasonable for those who know nothing about past observed effects. I expect that most parapsychologists would not opt for a prior symmetric around chance, but would still choose one with some mass below



chance. The final reason it is instructive to examine these Bayes factors is that they provide a quantitative challenge to skeptics to be explicit about their prior probabilities for the null and alternative hypotheses.

Dawson discussed the use of more complex Bayesian models for the analysis of the autoganzfeld data. She proposed a hierarchical model where the number of successes for each experiment followed a binomial distribution with hit rate  $p_i$ , and  $\text{logit}(p_i)$  came from a normal distribution with noninformative priors for the mean and variance. She then expanded this model to include heavier tails by allowing an additional scale parameter for each experiment. Her rationale for this expanded model was that there were clear outlier series in the data.

The hierarchical model proposed by Dawson is a reasonable place to start given only that there were several experiments trying to measure the same effect, conducted by different investigators. In the autoganzfeld database, the model could be expanded to incorporate the additional information available. Each experiment contained some sessions with static targets and some with dynamic targets, some sessions in which the sender and receiver were friends and others in which they were not and some information about the extraversion score of the receiver. All of this information could be included by defining the individual session as the unit of analysis, and including a vector of covariates for each session. It would then make sense to construct a logistic regression model with a component for each experiment, following the model proposed by Dawson, and a term  $X\beta$  to include the covariates. A prior distribution for  $\beta$  could include information from earlier ganzfeld studies. The advantage of using a Bayesian approach over a simple logistic regression is that information could be continually updated. Some of the recent work in Bayesian design could then be incorporated so that future trials make use of the best conditions.

Several of the discussants addressed the concept of replication. I agree with Mosteller's implication that it was unwise for the audience in my seminar to respond to my replication questions so quickly, and that was precisely my point. Most nonstatisticians do not seem to understand the complexity of the replication question. Parenthetically, when I posed the same scenario to an audience of statisticians, very few were willing to offer a quick opinion.

Bayarri and Berger provided an insightful discussion of the purpose of replication, offering quantitative answers to questions that were implicit in

my discussion. Their analyses suggest some alternatives to power analysis that might be considered when designing a new study to try to replicate a questionable result.

Morris addressed the question of what constitutes a replication of a meta-analysis. He distinguished between exact and conceptual replications. Using his distinction, the autoganzfeld meta-analysis could be viewed as a conceptual replication of the earlier ganzfeld meta-analysis. He noted that when such a conceptual replication offers results similar to those of the original meta-analysis, it lends legitimacy to the original results, as was the case with the autoganzfeld meta-analysis.

Greenhouse and Morris both noted the value of meta-analysis as a method of comparing different conditions, and I endorse that view. Conditions found to produce different effects in one meta-analysis could be explicitly studied in a conceptual replication. One of the intriguing results of the autoganzfeld experiments was that they supported the distinction between effect sizes for dynamic versus static targets found in the earlier ganzfeld work, and they supported the relationship between ESP and extraversion found in the meta-analysis by Honorton, Ferrari and Bem (1990).

Most modern parapsychologists, as indicated by Morris, recognize that demonstrating the validity of their preliminary findings will depend on identifying and utilizing "moderator variables" in future studies. The use of such variables will require more complicated statistical models than the simple binomial models used in the past. Further, models are needed for combining results from several different experiments, that don't oversimplify at the expense of lost information.

In conclusion, the anomalous effect that persists throughout the work reviewed in my paper will be better understood only after further experimentation that takes into account the complexity of the system. More realistic, and thus more complex, models will be needed to analyze the results of those experiments. This presents a challenge that I hope will be welcomed by the statistics community.

## ADDITIONAL REFERENCES

- ALLISON, P. (1979). Experimental parapsychology as a rejected science. *The Sociological Review Monograph* 27 271-291.
- BARBER, B. (1961). Resistance by scientists to scientific discovery. *Science* 134 596-602.
- BERGER, J. O. and DELAMPADY, M. (1987). Testing precise hypotheses (with discussion). *Statist. Sci.* 2 317-352.
- CHUNG, F. R. K., DIACONIS, P., GRAHAM, R. L. and MALLOWS, C. L. (1981). On the permanents of compliments of the direct sum of identity matrices. *Adv. Appl. Math.* 2 121-137.



- COCHRAN, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10** 101-129.
- COLLINS, H. and PINCH, T. (1979). The construction of the parnormal: Nothing unscientific is happening. *The Sociological Review Monograph* **27** 237-270.
- COLLINS, H. M. and PINCH, T. J. (1982). *Frames of Meaning: The Social Construction of Extraordinary Science*. Routledge & Kegan Paul, London.
- CORNFIELD, J. (1959). Principles of research. *American Journal of Mental Deficiency* **64** 240-252.
- DEMPSTER, A. P., SELWYN, M. R. and WEEKS, B. J. (1983). Combining historical and randomized controls for assessing trends in proportions. *J. Amer. Statist. Assoc.* **78** 221-227.
- DIACONIS, P. and GRAHAM, R. L. (1981). The analysis of sequential experiments with feedback to subjects. *Ann. Statist.* **9** 236-244.
- FISHER, R. A. (1932). *Statistical Methods for Research Workers*, 4th ed. Oliver and Boyd, London.
- FISHER, R. A. (1935). Has Mendel's work been rediscovered? *Ann. of Sci.* **1** 116-137.
- GALTON, F. (1901-2). Biometry. *Biometrika* **1** 7-10.
- GREENHOUSE, J., FROMM, D., IYENGAR, S., DEW, M. A., HOLLAND, A. and KASS, R. (1990). Case study: The effects of rehabilitation therapy for aphasia. In *The Future of Meta-Analysis* (K. W. Wachter and M. L. Straf, eds.) 31-32. Russell Sage Foundation, New York.
- GRIFFIN, D. (1988). Intuitive judgment and the evaluation of evidence. In *Enhancing Human Performance: Issues, Theories and Techniques Background Papers—Part I*. National Academy Press, Washington, D.C.
- HANSEN, G. (1990). Deception by subjects in psi research. *Journal of the American Society for Psychical Research* **84** 25-80.
- HUNTER, J. and SCHMIDT, F. (1990). *Methods of Meta-Analysis*. Sage, London.
- IYENGAR, S. and GREENHOUSE, J. (1988). Selection models and the file drawer problem (with discussion). *Statist. Sci.* **3** 109-135.
- LOUIS, T. A. (1984). Estimating an ensemble of parameters using Bayes and empirical Bayes methods. *J. Amer. Statist. Assoc.* **79** 393-398.
- MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22** 719-748.
- MORRIS, C. (1983). Parametric empirical Bayes inference: Theory and applications (rejoinder) *J. Amer. Statist. Assoc.* **78** 47-65.
- MORRIS, R. L. (1986). What psi is not: The necessity for experiments. In *Foundations of Parapsychology* (H. L. Edge, R. L. Morris, J. H. Rush and J. Palmer, eds.) 70-110. Routledge & Kegan Paul, London.
- MOSTELLER, F. and BUSH, R. R. (1954). Selected quantitative techniques. In *Handbook of Social Psychology* (G. Lindzey, ed.) 1 289-334. Addison-Wesley, Cambridge, Mass.
- MOSTELLER, F. and CHALMERS, T. (1991). Progress and problems in meta-analysis. *Statist. Sci.* To appear.
- OTERI, L., ed. (1975). *Quantum Physics and Parapsychology*. Parapsychology Foundation, New York.
- PINCH, T. J. and COLLINS, H. M. (1984). Private science and public knowledge: The Committee for the Scientific Investigation of Claims of the Paranormal and its use of the literature. *Social Studies of Science* **14** 521-546.
- PLATT, J. R. (1964). Strong inference. *Science* **146** 347-353.
- ROSENTHAL, R. (1966). *Experimenter Effects in Behavioral Research*. Appleton-Century-Crofts, New York.
- ROSENTHAL, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin* **86** 638-641.
- RYAN, L. M. and DEMPSTER, A. P. (1984). Weighted normal plots. Technical Report 394Z, Dana-Farber Cancer Inst., Boston, Mass.
- SAMANIEGO, F. J. and UTTS, J. (1983). Evaluating performance in continuous experiments with feedback to subjects. *Psychometrika* **48** 195-209.
- SMITH, M. and GLASS, G. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist* **32** 752-760.
- WACHTER, K. (1988). Disturbed by meta-analysis? *Science* **241** 1407-1408.
- WEST, M. (1985). Generalized linear models: Scale parameters, outlier accommodation and prior distributions. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.) 531-558. North-Holland Amsterdam.